AUDIO-COMPASS: REFERENCE-FREE COMPONENT-WISE EVALUATOR FOR AUDIO–TEXT SEMANTIC ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Many audio–text evaluation metrics summarize semantic alignment with a single overall score, motivating explicit assessment along consistent semantic dimensions. We introduce **Audio-Compass**, a 30B reference-free audio judge that produces overall and component-wise scores for sound events, sources, acoustic attributes, and temporal relations, along with structured explanations of matched, missing, and incorrect information. To train Audio-Compass, we construct **Compass-105k** with 105k fine-grained caption evaluations over 17.5k audio clips. Our perception–reasoning cascade consolidates descriptions from selected audio-language models into high-quality synthetic references, against which GPT-5-mini evaluates candidate captions to generate scores and explanations as training targets. Fine-tuned on this supervision, Audio-Compass evaluates audio–text alignment directly without references at inference time, supporting both audio-to-text caption evaluation and text-to-audio generation evaluation. Across BRACE, RELATE, PAM, and our new long-form benchmark BRACE-Long, Audio-Compass shows strong agreement with human judgments, achieving a Spearman correlation of 0.636 on PAM and a temporal-relation correlation of 0.370 on RELATE(OS), compared with up to 0.295 for the evaluated CLAP-based metrics. Further analysis suggests that holistic human judgments are more strongly associated with sound-event information than with acoustic attributes or temporal relations, motivating explicit component-wise evaluation.

1 INTRODUCTION

Audio-text semantic alignment is essential for a range of audio tasks, including audio captioning (Kim et al., 2019; Drossos et al., 2020), understanding (Gong et al., 2024; Chu et al., 2023), reasoning (Ghosh et al., 2024; 2025a), and generation (Liu et al., 2024; Majumder et al., 2024). The development of Large Audio Language Models (LALMs) (Ghosh et al., 2026; Yang et al., 2026) and omni-modal models (Xu et al., 2025; Google DeepMind, 2025) has substantially advanced audio–language modeling across these tasks. However, existing evaluation methods often reduce audio–text alignment to a single overall score, making it difficult to identify errors in specific semantic aspects such as sound events, sources, acoustic attributes, and temporal relations. Prior work further suggests that sound-event information has a disproportionately strong influence on overall caption assessment, while finer-grained aspects such as sources, acoustic attributes, and temporal relations may be under-reflected. This motivates component-wise evaluation that explicitly assesses each semantic aspect for more reliable and diagnostic assessment of audio–language models.

Despite this need, existing automatic evaluation metrics face several limitations in supporting component-wise assessment. Lexical metrics such as BLEU (Papineni et al., 2002) and CIDEr (Vedantam et al., 2015) measure n-gram overlap with references rather than semantic alignment with audio. Embedding-based scores such as CLAPScore (Xiao et al., 2024) reduce an entire description to a single similarity value. Recent methods leverage LALMs either as judges (Guo et al., 2025) or integrate them with CLAPScore (Lee et al., 2026), yet still produce only holistic scores. X-ACE (Wang et al., 2024) provides component-wise assessment but relies on reference captions, which are costly to collect at scale and unavailable when evaluating generated audio.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this paper, we present **Audio-Compass**, a 30B LALM reference-free evaluator for fine-grained audio–text alignment. Given an audio clip and a caption, Audio-Compass produces an *overall alignment score* with supporting reasoning, as well as component-wise scores for *sound events, sources, acoustic attributes, and temporal relations*. For each axis, Audio-Compass additionally provides structured reasoning that identifies *matched*, *missing*, and *incorrect* information with respect to the audio.

To train Audio-Compass, we construct **Compass-105k**, a large-scale dataset of 105k fine-grained caption evaluations over 17.5k audios. We leverage the strong evaluation capability of text-only LLMs—i.e., LLM-as-a-Judge—and replace the audio signal with high-quality reference captions as a textual proxy. Specifically, we first collect diverse captions generated from 13 LALMs. For reference construction, we select five models with strong audio-understanding capability based on MMAU performance. We then synthesize short, medium, and long references by consolidating the information consistently described across these captions, which suppresses model-specific hallucinations, and apply CLAP-based filtering to remove low-quality references. Finally, GPT-5-mini(Singh et al., 2025) evaluates each caption against its references along the four axes and assigns overall scores. Notably, the references are used only for data construction; Audio-Compass, initialized from Qwen3-Omni-30B-A3B-Instruct, is trained and evaluates captions directly from the audio.

We evaluate Audio-Compass on BRACE(Guo et al., 2025), RELATE(Kanamori et al., 2025), PAM(Deshmukh et al., 2024), and our newly constructed **BRACE-Long** benchmark which is human-annotated pairwise benchmark for long-form audio captions. Audio-Compass shows strong agreement with human judgments at both the overall and component levels, achieving an overall Spearman correlation of 0.636 on PAM and a sound-event correlation of 0.407 on RELATE(IS). On RELATE(OS), its temporal relation score reaches 0.370, outperforming CLAP-based metrics, which achieve up to 0.295. These results indicate that existing holistic metrics capture sound-event information relatively well, but are less effective at modeling fine-grained semantic structure such as event ordering.

Beyond benchmark performance, our analysis further confirms an event-centric tendency in human audio-caption judgments. Human ratings are strongly associated with sound-event information, while acoustic attributes and temporal relations have substantially weaker influence. This suggests that evaluations designed primarily to reproduce holistic human judgments may fail to adequately reflect certain semantic dimensions of audio–text alignment, further motivating component-wise evaluation that explicitly assesses these aspects. To support reproducibility and future research, we will publicly release Audio-Compass, Compass-105k, BRACE-Long and its annotations, together with the training and evaluation code.

2 Related Work

2.1 Audio-Text Semantic Alignment Metrics

Audio captioning is typically evaluated with *reference-based* metrics adapted from machine translation and image captioning, including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee & Lavie, 2005), CIDEr (Vedantam et al., 2015), and FENSE (Zhou et al., 2022). CLAIR-A(Wu et al., 2025) uses LLMs to provide holistic semantic scores and natural-language explanations. X-ACE (Wang et al., 2024) decomposes captions into four semantic factors—sound event, source, attribute, and relation—and evaluates each factor through graph-based matching against fine-grained references, enabling factor-wise assessment and more interpretable diagnosis of semantic errors. However, all of these approaches require reference captions.

Reference-free metrics instead directly measure audio–text correspondence. Contrastive Language–Audio Pretraining (CLAP)-based methods including MS-CLAP(Elizalde et al., 2023), LAION-CLAP(Wu et al., 2023), and M2D-CLAP(Niizumi et al., 2025) learn a shared embedding space for audio and text. CLAPScore (Xiao et al., 2024) uses pretrained CLAP representations to compute semantic similiarity between raw audio and text. Human-CLAP (Takano et al., 2025) fine-tunes pretrained LAION-CLAP using human subjective scores, thereby directly incorporating human judgments into audio–text alignment scoring. However, these metrics typically summarize audio–text alignment as a single similarity score, providing limited specific explanation of semantic errors.

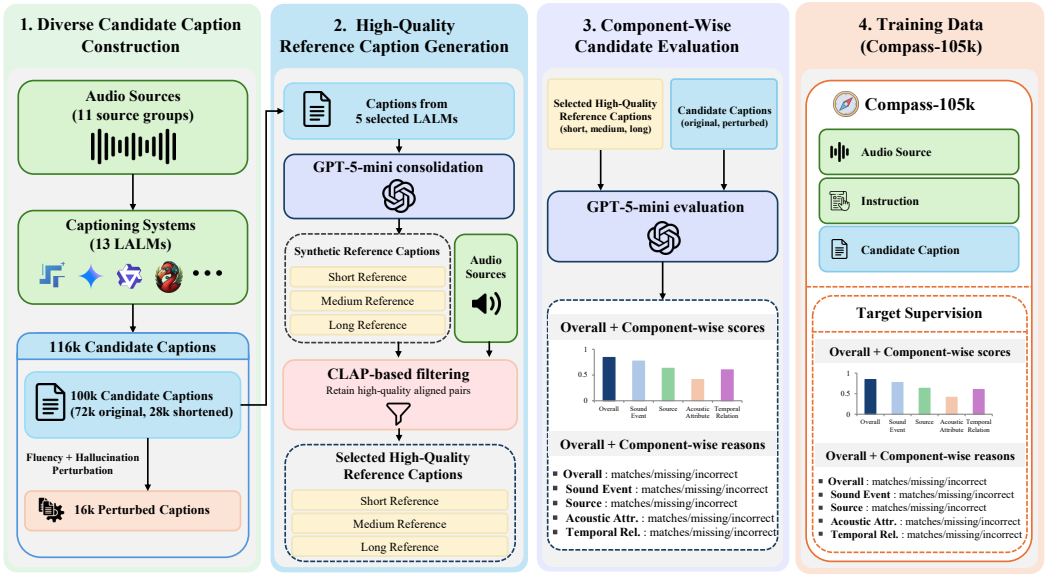


Figure 1: Compass-105k construction pipeline : We collect diverse candidate captions, consolidate selected LALM outputs into synthetic references, and use GPT-5-mini to generate fine-grained evaluation supervision. The resulting audio–caption pairs and supervision are used to train the reference-free judge.

2.2 Multimodal LLMs as Judges

Following LLM-as-a-Judge approaches (Zheng et al., 2023; Kim et al., 2024), multimodal large language models (MLLMs) have recently been explored as evaluators that directly assess the quality and grounding of generated text outputs with respect to multimodal inputs such as image (Lee et al., 2024a), video (Waheed et al., 2026). In the audio domain, AnyAudio-Judge (Li et al., 2026) is a dedicated 30B evaluator trained on a 105K-sample corpus that dynamically decomposes each instruction into multiple binary rubric items, evaluates them individually, and aggregates their satisfaction probabilities into an overall alignment score. On one hand, CAF-Score (Lee et al., 2026) uses LALMs for pointwise scoring, but raw discrete scores often produce ties. It therefore applies FLEUR-style score smoothing (Lee et al., 2024b) and combines the resulting score with CLAP, while still providing only a holistic evaluation. Consequently, existing LALM-based audio judges remain limited in identifying which semantic components are aligned or mismatched.

3 Method

We construct Compass-105k, which is the training supervision for Audio-Compass, using a perception–reasoning cascade pipeline. Motivated by prior findings that multimodal models can benefit from externalizing perceptual information into textual descriptions before downstream reasoning or evaluation (Weng et al., 2025; Ma et al., 2026), we decouple audio perception from fine-grained evaluation during dataset construction. The cascade pipeline first converts audio content into a rich textual representation and then leverages a strong text reasoning model to generate component-wise and overall evaluation supervision. We subsequently perform supervised fine-tuning on this teacher-generated supervision to obtain a reference-free audio judge.

3.1 Compass-105k

Diverse Audio–text Pool Construction. We collect audio from 11 source groups including AudioCaps (Kim et al., 2019) and Clotho-v2 (Drossos et al., 2020), and obtain candidate captions from 13 captioning systems from different model families and with varying audio-understanding capabilities, such as Qwen3-Omni(Xu et al., 2025) and Audio Flamingo 3(Ghosh et al., 2025a). The lists of

audio sources and captioning systems are described in Appendix A.1 and A.1.1, respectively. Since models with different perceptual capabilities can produce captions with varying levels of correctness and completeness, we intentionally include diverse captioning systems to cover a broad range of caption quality. To further increase diversity in the candidate pool and reduce reliance on surface-level cues such as caption length, we generate long-to-short variants from five systems while preserving key semantic content, and introduce hallucination and fluency perturbations. The prompts used for data construction are provided in Appendix A.3.

Synthetic High-Quality Reference Caption Generation. Some audio sources do not provide human-written reference captions, and even when references are available, human-written captions are often sparse and centered on salient sound events, and often omit details such as acoustic attributes, secondary events, and temporal relations. We therefore construct synthetic high-quality reference set to provide richer and more consistent evaluation references. For each audio, we obtain captions from five audio-language models selected using MMAU performance (Sakshi et al., 2025) as a proxy for general audio-understanding capability, and use GPT-5-mini(Singh et al., 2025) to consolidate their descriptions. From the consolidated audio information, we generate short, medium, and long references variants. Then we apply CLAP-based quality filtering using the minimum CLAPScore across the short, medium, and long reference variants for each audio, removing the lowest-scoring 20% within each source group and retaining 19.5k of 24k audios.

Component-wise Candidate Evaluation. Given the synthetic references, we formulate candidate evaluation as a text-based reasoning task. GPT-5-mini evaluates each candidate caption against the references set and generates scores and explanations for four semantic components—sound event, source, acoustic attribute, and temporal relation—as well as an overall evaluation.

Compass-105k statistics We combine 100k original or shortened candidates with 16k perturbed candidates, resulting in a dataset of 116k samples across 19.5k audios. We split the dataset by audio ID, using 17.5k audios comprising 105k samples for training. The validation and test sets each contain approximately 1k audios and 5.5k samples. Detailed dataset statistics are provided in Appendix A.1.

3.2 AUDIO-COMPASS

We fine-tune Qwen3-Omni-30B-A3B-Instruct with Compass-105k; that is we distill the reference-based text evaluation capability of GPT-5-mini into the LALM without the synthetic references but audio sources.

Given $\mathcal{D} = \{(a_i, x_i, y_i, t_i)\}_{i=1}^{N}$, where a_i denotes the audio input, x_i the evaluation instruction, y_i the candidate caption, and t_i the target supervision containing overall and component-wise scores and explanations. The model is trained to autoregressively generate the target supervision t_i conditioned on the audio, evaluation instruction, and candidate caption using the standard negative log-likelihood objective:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{|t_i|}\log P_\theta\left(t_{i,j} \mid t_{i,<j}, a_i, x_i, y_i\right).$$

At inference time, Audio-Compass directly generates overall and component-wise scores and explanations from given an audio and candidate caption without synthetic references or the teacher model, using the evaluation prompt shown in Figure 16.

4 EXPERIMENTS

4.1 BASELINES

CLAP-based Metric. We include MS-CLAP, LAION-CLAP, M2D-CLAP and Human-CLAP (Elizalde et al., 2023; Wu et al., 2023; Niizumi et al., 2025; Takano et al., 2025) as reference-free embedding-based baselines, which measure audio–text alignment using the similarity between audio and text representations.

LALM-as-a-Judge. As for LALM-as-a-judges, we include Qwen3-Omni-Instruct, Gemini-3-Flash, Gemini-3.1-Pro (Xu et al., 2025; Google DeepMind, 2025), and AnyAudio-Judge (Li et al., 2026). For these baselines, we follow the evaluation prompt (Figure A.5) used in FLEUR (Lee et al., 2024b), and report raw overall scores. For open-weight models, we additionally apply FLEUR score smoothing using token-level probabilities. Specifically, letting $p(i, j)$ denote the probability of generating digit $i \in \{0, \ldots, 9\}$ at the j-th decimal place, the smoothed score is computed as

$$S_{\mathrm{FLEUR}} = \sum_{j=1}^{2} 10^{-j} \sum_{i=0}^{9} i\, p(i, j).$$

Ensemble of CLAPScore and LALM We compare CAF-Score (Lee et al., 2026), which linearly combines CLAPScore with the FLEUR-smoothed LALM score: $S_{\mathrm{CAF}} = \alpha S_{\mathrm{CLAP}} + (1 - \alpha) S_{\mathrm{FLEUR}}$, where hyperparameter α is set as 0.8.

Reference-based Metrics. For benchmarks where human-written reference captions are available such as BRACE and BRACE-Long, we additionally compare against FENSE and CLAIR-A (Zhou et al., 2022; Wu et al., 2025). These methods evaluate candidate captions by comparing them with reference captions. Note that these methods are not directly applicable to the text-to-audio (TTA) meta-evaluation benchmarks, such as RELATE(Kanamori et al., 2025) and PAM(Deshmukh et al., 2024) In these benchmarks, the evaluation target is the generated audio itself, leading to a lack of reference audio.

4.2 Evaluation Benchmarks

BRACE. BRACE(Guo et al., 2025) evaluates audio-caption ranking under two settings. BRACE-Main contains human-annotated preferences over human–human(HH), human–machine(HM), and machine–machine(MM) caption pairs. BRACE-Hallucination pairs an original human caption with a synthetically hallucinated version and tests whether an evaluator assigns a higher score to the original caption. For each pair, the evaluator ranks the two captions based on their scores, and pairwise accuracy is measured by whether its ranking agrees with the benchmark preference. Most captions in BRACE are relatively short, typically within 30 words, motivating our evaluation on longer captions with BRACE-Long.

BRACE-Long. Recent LALMs increasingly generate long and detailed audio descriptions containing multiple sound events, attributes, and temporal information, which are not well represented by existing pairwise benchmarks. To evaluate such outputs, we construct **BRACE-Long**, a human-annotated pairwise benchmark for long-form audio captions. Starting from 575 audios from BRACE-Main, we collect captions generated by eight captioning systems and construct cross-system caption pairs. Six annotators evaluate 150 caption pairs from 150 audio clips (300 captions), providing preferences and written rationales. After excluding three pairs with repetitive captions and 18 ties, 129 pairs remain. For the more detailed benchmark construction, please refer Appendix A.2.

RELATE. RELATE(Kanamori et al., 2025) is a subjective evaluation dataset for audio–text relevance containing both natural and synthesized audio. It provides human scores for overall relevance (REL), sound-event inclusion (IS), and sound-event ordering (OS), which we use to measure the agreement between evaluator scores and human judgments.

PAM. We use the text-to-audio human listening-test data released with PAM(Deshmukh et al., 2024), which contains both original AudioCaps audio and audio generated by multiple TTA systems. The dataset provides human scores for overall quality (OVL) and audio–text relevance (REL); we use the REL scores to evaluate how well automatic evaluator scores align with human judgments of audio–text relevance.

5 Results

5.1 Results on BRACE and BRACE-Long

Table 1: Pairwise accuracy(%) on BRACE-Main and BRACE-Hallucination. Model score ties count as incorrect.

Model (Method)	BRACE-Main									BRACE-Hallucination		
	AudioCaps-Main				Clotho-Main				Overall	AudioCaps -Hallu	Clotho -Hallu	Overall
	HH	HM	MM	Total	HH	HM	MM	Total				
Reference-free												
CLAP-based scoring												
MS-CLAP	64.03	48.76	55.08	53.71	67.07	70.92	65.93	67.95	61.42	72.99	85.23	80.61
LAION-CLAP	63.31	77.30	65.42	69.78	59.28	75.64	65.63	68.62	69.15	87.03	84.31	85.34
M2D-CLAP	65.47	80.22	63.99	70.48	67.66	66.21	62.81	64.69	67.35	91.03	83.93	86.61
Human-CLAP	59.71	71.24	64.35	66.46	56.89	73.08	65.33	67.21	66.87	85.80	83.09	84.11
Raw Score (LALM-as-a-Judge)												
Qwen3-Omni-Instruct	16.55	48.54	41.00	40.96	35.93	54.62	45.04	47.52	44.51	81.78	82.47	82.21
Gemini-3-Flash	31.65	60.90	55.97	54.93	41.92	65.42	48.74	54.18	54.53	91.29	91.81	91.61
Gemini-3.1-Pro	13.67	50.11	46.17	43.76	14.97	50.10	45.63	43.52	43.63	91.40	87.65	89.07
AnyAudio-Judge	42.45	64.27	60.25	59.65	56.89	66.01	57.93	60.84	60.30	91.90	84.91	87.55
FLEUR Score (LALM)												
Qwen3-Omni-Instruct	53.24	78.65	66.31	69.52	67.66	74.66	61.63	67.28	68.31	**98.48**	**98.00**	**98.18**
CAF-Score (CLAP + LALM)												
Qwen3-Omni-Instruct+MS-CLAP	64.03	60.67	63.99	62.71	67.66	**79.17**	**70.07**	**73.21**	68.39	93.52	96.61	95.44
Qwen3-Omni-Instruct+M2D-CLAP	**67.63**	**86.97**	**67.91**	**75.28**	**70.66**	77.80	66.52	71.28	**73.11**	97.96	96.59	97.11
Audio-Compass (Ours)												
Audio-Compass + raw	53.24	79.33	60.78	67.07	65.87	63.85	61.19	62.77	64.74	93.59	89.00	90.74
Audio-Compass + FLEUR	58.99	82.92	67.56	72.49	70.06	70.92	69.48	70.10	71.19	97.54	96.10	96.65
Reference-based												
FENSE	61.15	84.49	67.38	73.28	**56.89**	84.68	64.15	70.98	72.04	**96.76**	**96.18**	**96.40**
CLAIR-A	**66.19**	**90.79**	**67.74**	**76.51**	56.29	**91.94**	**65.63**	**74.39**	**75.36**	91.45	91.55	91.51

BRACE-Main Results. On BRACE-Main (Table 1), Audio-Compass achieves the highest overall pairwise accuracy among the raw-score LALM judges, outperforming both proprietary Gemini-series models. Applying FLEUR score smoothing further improves the accuracy to 71.19%, yielding the strongest performance among the single-model reference-free evaluators. CAF-Score achieves a higher accuracy by combining FLEUR-smoothed LALM scores with CLAP-based similarity through a weighted metric ensemble, leveraging complementary evaluation signals. In contrast, Audio-Compass achieves its performance using a single learned evaluator, suggesting that its stronger standalone scoring could provide further headroom when combined with complementary signals through metric-level fusion in future work.

Table 2: Pairwise accuracy (%, ↑) on BRACE-Long. Model score ties count as incorrect.

Model	Raw ↑	FLEUR ↑
Reference-free		
Qwen3-Omni-Instruct	42.64	**73.64**
Gemini-3-Flash	62.79	–
Gemini-3.1-Pro	66.67	–
Audio-Compass (Ours)	**68.99**	68.99
Reference-based		
CLAIRA	62.79	–

BRACE-Hallucination Results. On BRACE-Hallucination (Table 1), Audio-Compass achieves an overall pairwise accuracy of 90.74% using the raw score, showing competitive performance among LALM-based judges. Applying FLEUR score smoothing further improves the accuracy to 96.65%, outperforming both proprietary Gemini-series judges. This performance is also competitive with CAF-Score, despite Audio-Compass relying on a single learned evaluator rather than combining complementary CLAP- and LALM-based signals. However, Qwen3-Omni-Instruct with FLEUR achieves the highest accuracy of 98.18%, indicating that the benefits of score smoothing and metric fusion can vary across evaluation settings.

BRACE-Long Results. On BRACE-Long (Table 2), Audio-Compass achieves the highest raw pairwise accuracy of 68.99%, outperforming the proprietary Gemini-3-Flash and Gemini-3.1-Pro judges as well as the reference-based CLAIR-A evaluator (62.79%). Notably, Audio-Compass achieves this performance without requiring reference captions at inference time, indicating its effectiveness in distinguishing human-preferred captions even for long and detailed descriptions. When applying FLEUR score smoothing, the accuracy of Audio-Compass remains at 68.99%. In con-

trast, Qwen3-Omni-Instruct improves substantially from 42.64% to 73.64%, achieving the highest FLEUR-based accuracy on BRACE-Long.

5.2 Results on RELATE and PAM

5.2.1 Overall Relevance

Table 3: Correlation with human relevance judgments on RELATE and PAM. AnyAudio-Judge uses rubric-based scores; other LALMs use raw scores; CLAP-based methods use similarity scores.

Model / component	RELATE (REL)			PAM		
	LCC	SRCC	KTAU	LCC	SRCC	KTAU
CLAP-Based						
MS-CLAP	0.403	0.380	0.264	0.345	0.312	0.215
LAION-CLAP	0.405	0.381	0.264	0.461	0.452	0.314
M2D-CLAP	<u>0.492</u>	0.457	0.322	0.566	0.548	0.388
Human-CLAP	0.490	<u>0.463</u>	0.326	0.576	0.557	0.396
LALM-as-a-Judge						
Qwen3-Omni-Instruct	0.374	0.432	<u>0.343</u>	0.420	0.384	0.299
Gemini-3-Flash	0.280	0.293	0.229	0.276	0.369	0.286
Gemini-3.1-Pro	0.417	0.377	0.306	0.442	0.408	0.327
AnyAudio-Judge	0.392	0.381	0.264	0.614	0.601	0.435
CAF-Score (CLAP + LALM)						
Qwen3+MS-CLAP	0.372	0.329	0.227	0.521	0.483	0.340
Qwen3+M2D-CLAP	**0.526**	**0.520**	**0.367**	0.608	<u>0.606</u>	0.436
Audio-Compass (Ours)						
Overall	0.449	0.428	0.309	**0.642**	**0.636**	**0.475**
Sound event	0.460	0.450	0.332	<u>0.618</u>	0.603	<u>0.454</u>

Table 3 reports correlations with human relevance judgments on RELATE(REL) and PAM. On RELATE(REL), Qwen3+M2D-CLAP achieves the highest correlation with human judgments across all three metrics, demonstrating the effectiveness of combining complementary LALM and CLAP-based evaluation signals. Among the outputs of Audio-Compass, the Sound event score correlates more strongly with human judgments than the Overall score, suggesting that overall relevance judgments in RELATE are strongly influenced by sound-event alignment. We further examine this tendency across benchmarks in Sec. 6.1.

On PAM, however, Audio-Compass achieves the highest correlation with human relevance judgments across all three metrics, surpassing not only individual LALM and CLAP-based evaluators but also the CAF-Score metric ensembles. This indicates that strong human alignment can be achieved by a single learned evaluator without explicitly combining multiple evaluation signals. Together, these results suggest that metric-level fusion can provide substantial benefits when complementary signals are available, while improving the standalone evaluator itself remains an important foundation and may provide additional headroom for future metric fusion.

5.2.2 RELATE(IS) and RELATE(OS) Results.

Table 4 reports correlations with human judgments of sound-event inclusion (IS) and ordering (OS) on RELATE. On RELATE(IS), the Sound event score of Audio-Compass shows the highest correlation with human judgments. Human-CLAP also performs strongly on IS despite being trained only on the RELATE(REL) training set. This suggests that overall relevance judgments in RELATE are closely related to sound-event inclusion. In contrast, Human-CLAP performs worse on OS, suggesting that temporal ordering is less reflected in overall relevance. On RELATE(OS), the Overall score of Audio-Compass shows the highest correlation, followed by the Temporal relation and Sound event scores. This suggests that even when evaluating event ordering, human judgments may still be influenced by broader sound-event alignment. We further analyze this tendency in Sec. 6.1.

Table 4: Correlation with human judgments on RELATE inclusion of sound event and order of sound event. AnyAudio-Judge uses rubric-based scores; LALM-as-a-Judge rows use raw scores; CLAP-based methods use similarity scores.

Model / component	RELATE (IS)			RELATE (OS)		
	LCC	SRCC	KTAU	LCC	SRCC	KTAU
CLAP-Based						
MS-CLAP	0.359	0.341	0.240	0.285	0.295	0.207
LAION-CLAP	0.368	0.356	0.248	0.259	0.267	0.183
Human-CLAP	**0.430**	**0.423**	0.300	0.260	0.267	0.185
LALM-as-a-Judge						
Qwen3-Omni-Instruct	0.355	0.396	**0.318**	0.320	0.367	0.292
Gemini-3-Flash	0.248	0.302	0.239	0.216	0.250	0.195
Gemini-3.1-Pro	0.381	0.365	0.300	0.390	0.395	**0.313**
AnyAudio-Judge	0.396	0.398	0.281	0.352	0.365	0.257
Audio-Compass (Ours)						
Overall	0.389	0.378	0.277	**0.402**	**0.404**	0.291
Sound event	0.406	0.407	0.304	0.348	0.358	0.264
Temporal relation	0.164	0.133	0.099	0.367	0.370	0.270

6 Discussion

6.1 What Do Human Audio–Caption Judgments Emphasize?

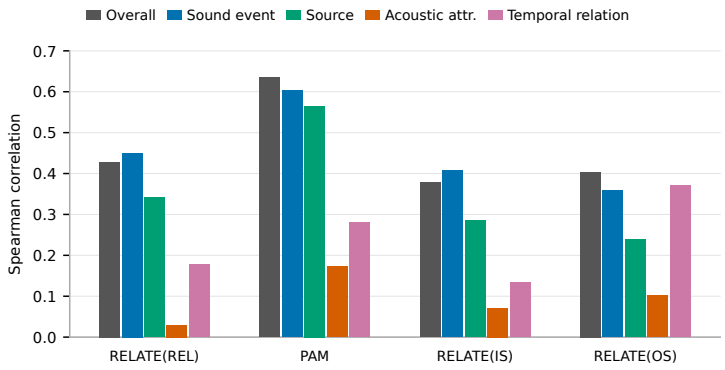


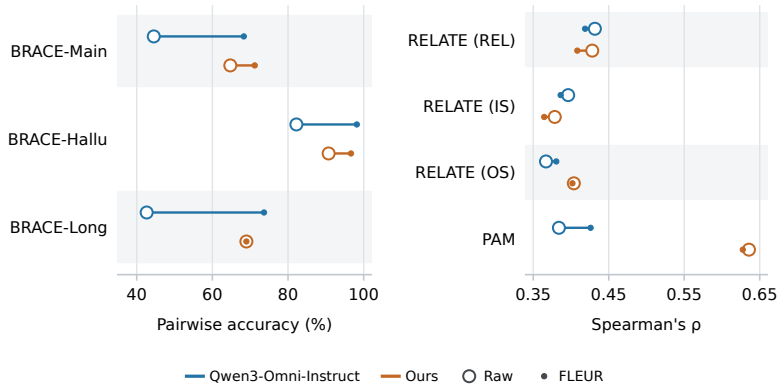
Figure 2: Spearman correlations between Audio-Compass outputs and mean human ratings across benchmarks. The component with the highest correlation varies across evaluation benchmarks.

Analysis setup. Each audio–caption sample has a mean human rating and five scores from Audio-Compass: Overall, Sound event, Source, Acoustic attribute, and Temporal relation. Within each benchmark, we compute Spearman's rank correlation between the mean human ratings and each model score across matched samples. A higher correlation indicates that the ranking induced by a model score more closely agrees with human judgments. By comparing these correlations across axes and benchmarks, we examine which semantic aspects are most strongly associated with human judgments and how these associations vary across evaluation tasks.

Event-centric tendency in human judgments. Across benchmarks, human judgments are strongly aligned with sound-event information. In RELATE(REL), PAM, and RELATE(IS), Sound event shows the strongest correlation among the component scores, while Acoustic attribute and Temporal relation are substantially weaker. This suggests that human evaluation of short captions is largely driven by whether the salient sound events and their sources are correctly captured. This tendency persists even in RELATE(OS): despite explicitly asking annotators to judge temporal order, Sound event remains comparably correlated with human judgments to Temporal relation, while the holistic Overall score correlates most strongly. Taken together, these results suggest that human judgments of audio–caption alignment exhibit an event-centric tendency, while other semantic

aspects, such as acoustic attributes and temporal relations, may be underrepresented in holistic judgments. Therefore, future evaluation methods should go beyond simply reproducing this tendency and explicitly capture semantic aspects that may be underweighted in holistic human evaluation.

6.2 Score Resolution and the Limits of FLEUR Smoothing



(a) (b)

Figure 3: Comparison of raw and FLEUR-smoothed scores. (a) Pairwise evaluation measured by accuracy. (b) Pointwise evaluation measured by Spearman correlation. Open and filled markers denote raw and fleur scores. Pairwise ties count as incorrect.

Pairwise gains through tie resolution. FLEUR uses score-token probabilities to produce effectively continuous scores with substantially fewer ties (Lee et al., 2024b). As shown in Figure 3(a), Qwen3-Omni-Instruct improves across all three pairwise benchmarks, while Audio-Compass improves on BRACE-Main and BRACE-Hallucination but remains unchanged on BRACE-Long.

Because ties are counted as incorrect, smoothing can improve pairwise accuracy simply by resolving them. Moreover, over 96% of pairs correctly ordered by the raw scores remain correctly ordered after smoothing across all model–benchmark settings. Thus, the gains largely come from increased score resolution while preserving existing correct rankings.

Finer scores do not always improve rank correlation. Pointwise evaluation shows a different pattern (Figure 3(b)). For Qwen3-Omni-Instruct, Spearman correlation increases on PAM and RELATE(OS) but decreases on RELATE(REL) and RELATE(IS). For Audio-Compass, it decreases across all four benchmarks.

The key difference is that pairwise evaluation can benefit directly from resolving ties, whereas Spearman correlation measures how well the ordering of all samples matches the human ranking. Because FLEUR produces finer-grained scores, it can change the relative ordering among samples. These changes improve correlation only when the new ordering is more consistent with human ratings; otherwise, correlation can decrease. Thus, finer score resolution does not necessarily lead to better agreement with human judgments. These results suggest that further research is needed to understand how score resolution can better capture meaningful differences in caption quality.

7 Conclusion

We introduced **Audio-Compass**, a reference-free audio judge for component-wise and overall caption evaluation, trained on **Compass-105k**, comprising 105k fine-grained evaluations over 17.5k audio clips. Using a perception–reasoning cascade, Audio-Compass distills reference-based supervision into a judge that evaluates captions directly from audio without references at inference time. Together, Audio-Compass and Compass-105k provide a foundation for audio-grounded caption evaluation that goes beyond a single overall score.

AI USE STATEMENT

Generative AI tools were used throughout this work for research assistance, dataset construction, software development, and manuscript preparation. For dataset construction, we used outputs from 13 audio-language captioning systems to construct a diverse candidate-caption pool. GPT-5-mini was further used to consolidate selected model outputs into synthetic anchor captions, generate data augmentations, and produce component-wise evaluation supervision, as described in Sec. 3. We also used OpenAI Codex to assist with code generation and debugging. In addition, ChatGPT were used to support literature retrieval and discovery, research ideation and execution, drafting portions of the manuscript, and improving the clarity and readability of the writing. The authors reviewed the resulting research artifacts and take responsibility for the final content of the paper.

ETHICS STATEMENT

This work involves human annotation for the construction of BRACE-Long. Six annotators were recruited through a university community and evaluated 160 audio–caption pairs, including 150 benchmark pairs and 10 quality-control pairs, following detailed guidelines. Annotators were not shown which model generated each caption or whether an item was included for quality control. Each annotator received KRW 75,000 as compensation. Personal information required for recruitment and administrative processing was collected separately and was not included in the annotation dataset or research analysis.

Our training data were constructed using outputs from multiple audio-language models and GPT-5-mini. These synthetic annotations may inherit errors or biases from the models used during data construction. We therefore treat them as training supervision rather than human-verified ground truth and evaluate our model against independent human judgments.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the dataset construction pipeline, model training procedure, evaluation protocols, prompts, and experimental settings in the main paper and appendix. To further support reproducibility and future research, we plan to publicly release Audio-Compass, the Compass-105k training dataset, the BRACE-Long benchmark and its annotations, as well as the training and evaluation code after the conclusion of the ICLR 2027 review process.

REFERENCES

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

Soham Deshmukh, Benjamin Elizalde, and Huaming Wang. Audio Retrieval with WavText5K and CLAP Training. In *Interspeech 2023*, pp. 2948–2952, 2023. doi: 10.21437/Interspeech.2023-1136.

Soham Deshmukh, Dareen Alharthi, Benjamin Elizalde, Hannes Gamper, Mahmoud Al Ismail, Rita Singh, Bhiksha Raj, and Huaming Wang. Pam: Prompting audio-language models for audio quality assessment. *arXiv preprint arXiv:2402.00282*, 2024.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740, 2020. doi: 10.1109/ICASSP40776.2020.9052990.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2022.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6288–6313, 2024.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 41819–41886. Curran Associates, Inc., 2025a. doi: 10.52202/085713-1396. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3babb6b453cb59d87cb58a1219ef914b-Paper-Conference.pdf.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. *arXiv preprint arXiv:2503.03983*, 2025b.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, Sang-gil Lee, Jaehyeon Kim, Aya Aljafari, et al. Audio flamingo next: Next-generation open audio-language models for speech, sound, and music. *arXiv preprint arXiv:2604.10905*, 2026.

Yuan Gong, Hongyin Luo, Alexander Liu, Leonid Karlinsky, and James R Glass. Listen, think, and understand. In *International Conference on Learning Representations*, volume 2024, pp. 18516–18545, 2024.

Google DeepMind. Gemini 3 flash model card. https://deepmind.google/models/model-cards/gemini-3-flash/, 2025. Published December 17, 2025.

Tianyu Guo, Hongyu Chen, Hao Liang, Meiyi Qiang, Bohan Zeng, Linzhuang Sun, Bin CUI, and Wentao Zhang. Brace: A benchmark for robust audio caption quality evaluation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-1366. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3a3ec21884760f891752e0661ecc0b0b-Paper-Datasets_and_Benchmarks_Track.pdf.

Il-Young Jeong and Jeongsoo Park. Cochlscene: Acquisition of acoustic scene data using crowdsourcing. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 17–21, 2022. doi: 10.23919/APSIPAASC55919.2022.9979822.

Yusuke Kanamori, Yuki Okamoto, Taisei Takano, Shinnosuke Takamichi, Yuki Saito, and Hiroshi Saruwatari. RELATE: Subjective evaluation dataset for automatic evaluation of relevance between text and audio . In *Interspeech 2025*, pp. 3155–3159, 2025. doi: 10.21437/Interspeech.2025-1830.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, 2019.

Seungone Kim, Jay Shin, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Ryan Shin, Sungdong Kim, James Thorne, Minjoon Seo, et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations*, volume 2024, pp. 29927–29962, 2024.

Insung Lee, Taeyoung Jeong, Haejun Yoo, Du-Seong Chang, and Myoung-Wan Koo. Caf-score: Calibrating clap with lalms for reference-free audio captioning evaluation, 2026. URL https://arxiv.org/abs/2603.19615.

Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11286–11315, 2024a.

Yebin Lee, Imseong Park, and Myungjoo Kang. Fleur: An explainable reference-free evaluation metric for image captioning using a large multimodal model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3732–3746, 2024b.

Haitao Li, Tian Tan, Yuguang Yang, Shan Yang, and Xie Chen. Anyaudio-judge: A dynamic rubric-based benchmark and evaluator for audio instruction following, 2026. URL https://arxiv.org/abs/2606.03116.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024.

Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, et al. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception. In *International Conference on Learning Representations*, volume 2026, pp. 132277–132304, 2026.

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 564–572, 2024.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354, 2024. doi: 10.1109/TASLP.2024.3419446.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. A multi-device dataset for urban acoustic scene classification. In *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2018. URL https://api.semanticscholar.org/CorpusID:50785886.

Irene Martin Morato and Annamaria Mesaros. Diversity and bias in audio captioning datasets. In *Detection and Classication of Acoustic Scenes and Events*, pp. 90–94. DCASE, 2021.

Daisuke Niizumi, Daiki Takeuchi, Masahiro Yasuda, Binh Thien Nguyen, Yasunori Ohishi, and Noboru Harada. M2d-clap: Exploring general-purpose audio-language representations beyond clap. *IEEE Access*, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pp. 1015–1018. ACM Press, 2015. ISBN 978-1-4503-3459-4. doi: 10.1145/2733373.2806390. URL http://dl.acm.org/citation.cfm?doid=2733373.2806390.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 84929–84964, 2025.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 1041–1044, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450330633. doi: 10.1145/2647868.2655045. URL https://doi.org/10.1145/2647868.2655045.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Taisei Takano, Yuki Okamoto, Yusuke Kanamori, Yuki Saito, Ryotaro Nagase, and Hiroshi Saruwatari. Human-clap: Human-perception-based contrastive language-audio pretraining. In *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 131–136. IEEE, 2025.

Gemma Team. Gemma 3n, 2025. URL https://ai.google.dev/gemma/docs/gemma-3n.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, 2015. doi: 10.1109/CVPR.2015.7299087.

Abdul Waheed, Zhen Wu, Dareen Alharthi, Seungone Kim, and Bhiksha Raj. Videojudge: Bootstrapping enables scalable supervision of mllm-as-a-judge for video understanding. In *International Conference on Learning Representations*, volume 2026, pp. 103923–103955, 2026.

Qian Wang, Jia-Chen Gu, and Zhen-Hua Ling. X-ace: Explainable and multi-factor audio captioning evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12273–12287, 2024.

Zi-Han Weng, Lucas Gomez, Taylor Webb, and Pouya Bashivan. Caption this, reason that: Vlms caught in the middle. *ArXiv*, abs/2505.21538, 2025. URL https://api.semanticscholar.org/CorpusID:278959992.

Tsung-Han Wu, Joseph E Gonzalez, Trevor Darrell, and David M Chan. Clair a: Leveraging large language models to judge audio captions. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7. IEEE, 2025.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Feiyang Xiao, Jian Guan, Qiaoxi Zhu, Xubo Liu, Wenbo Wang, Shuhan Qi, Kejia Zhang, Jianyuan Sun, and Wenwu Wang. A reference-free metric for language-queried audio source separation using contrastive language-audio pretraining. *arXiv preprint arXiv:2407.04936*, 2024.

LLM-Core-Team Xiaomi. Mimo-audio: Audio language models are few-shot learners, 2025. URL https://github.com/XiaomiMiMo/MiMo-Audio.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Chen Yang, Chufan Yu, Hanfu Chen, Jie Zhu, Jingqi Chen, Ke Chen, Wenxuan Wang, Yang Wang, Yaozhou Jiang, Yi Jiang, et al. Moss-audio technical report. *arXiv preprint arXiv:2606.01802*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Zelin Zhou, Zhiling Zhang, Xuenan Xu, Zeyu Xie, Mengyue Wu, and Kenny Q. Zhu. Can audio captions be evaluated with image caption metrics? In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 981–985, 2022. doi: 10.1109/ICASSP43922.2022.9746427.

A Appendix

A.1 Dataset Statistics

A.1.1 Audio Sources

Table 5 summarizes 11 sources: AudioCaps (Kim et al., 2019), Clotho-v2 (Drossos et al., 2020), MACS (Morato & Mesaros, 2021), WavCaps (Mei et al., 2024), WavText5K (Deshmukh et al., 2023), ESC-50 (Piczak, 2015), FSD50K (Fonseca et al., 2022), UrbanSound8K (Salamon et al., 2014), CochlScene (Jeong & Park, 2022), TUT Urban (Mesaros et al., 2018), and BBC Sound Effects. The final dataset contains 19,484 unique audios, split by audio ID into 17,536 training and 974 each validation and test audios. The corresponding sample counts are 104,401, 5,813, and 5,763.

Table 5: Unique audios by source in the final dataset. All is the sum of Train, Valid (validation), and Test. Splits are disjoint by audio ID; each audio ID is counted once.

Audio source	Unique audios			
	All	Train	Valid	Test
AudioCaps	1,919	1,721	98	100
BBC Sound Effects	1,769	1,588	87	94
CochlScene	1,525	1,370	82	73
ESC-50	1,564	1,407	75	82
FSD50K	1,525	1,373	66	86
MACS	1,906	1,717	96	93
TUT Urban	1,885	1,710	86	89
UrbanSound8K	1,938	1,749	98	91
WavCaps	1,928	1,721	113	94
WavText5K	1,973	1,787	95	91
Clotho-v2	1,552	1,393	78	81
Total	**19,484**	**17,536**	**974**	**974**

Figure 4 counts caption lengths per retained sample and original audio-file durations per unique audio ID. Most captions are short (65.3% have fewer than 40 words), with a long tail pooled into the 200+ bin (9.8%). Most audios are at most 10 seconds long (71.3%), while 7.6% exceed 60 seconds.

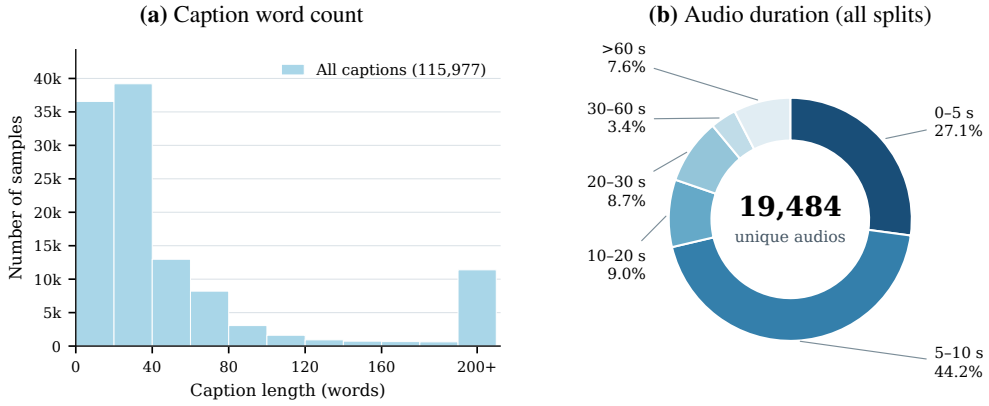


Figure 4: Dataset statistics across all splits. (a) Word count distribution of the 115,977 candidate captions. (b) Duration distribution of the 19,484 unique audio files.

A.1.2 USED LALMS

Table 6 lists 13 captioning systems, including Audio Flamingo (Ghosh et al., 2025b;a; 2026), GAMA (Ghosh et al., 2024), LTU (Gong et al., 2024), MiMo-Audio (Xiaomi, 2025), and Gemma (Team, 2025; 2026). The retained pool comprises 71,960 original captions, 27,773 shortened variants from five systems, and 16,244 GPT-5-mini perturbations (12,632 hallucination and 3,612 fluency perturbations). Shortened variants retain their source-system attribution; perturbations are counted separately.

Table 6: Caption counts by captioning system after filtering. Original and Shortened refer to all splits; All is their sum and is split into Train, Valid (validation), and Test. Shortened captions are attributed to their original captioning system. GPT-5-mini perturbations are counted separately.

Captioning system	Original	Shortened	All	Train	Valid	Test
AF2	5,556	0	5,556	4,997	277	282
AF3	5,555	0	5,555	5,003	285	267
AF-Next-Captioner	5,549	5,554	11,103	9,974	584	545
AF-Next	5,555	0	5,555	5,000	279	276
GAMA	5,556	0	5,556	5,010	268	278
Gemini-3-Flash	5,555	5,555	11,110	10,000	543	567
Gemma3n	5,552	0	5,552	5,012	285	255
Gemma4-12B-IT	5,540	0	5,540	4,979	277	284
LTU	5,552	0	5,552	4,982	282	288
MiMo-Audio	5,543	5,555	11,098	10,006	530	562
MOSS-Audio-8B-Thinking	5,537	0	5,537	4,985	279	273
Qwen3-Captioner	5,530	5,555	11,085	9,940	555	590
Step-Audio-R1.1	5,380	5,554	10,934	9,883	535	516
Original/shortened subtotal	71,960	27,773	99,733	89,771	4,979	4,983
Perturbed candidates	–	–	16,244	14,630	834	780
Total	–	–	**115,977**	**104,401**	**5,813**	**5,763**

A.2 BRACE-LONG STATISTICS

Figure 5 summarizes the final BRACE-Long pairwise evaluation set used in our experiments: 129 caption pairs with at least 4:2 agreement, comprising 258 captions and 129 unique audios (60 from AudioCaps and 69 from Clotho). The statistics exclude dummy quality-control pairs and 3:3 vote ties. Caption lengths have a mean of 314.0 words and a median of 216.5 words, ranging from 77 to 915 words. Original audio durations have a mean of 16.83 seconds and a median of 16.08 seconds, ranging from 7.23 to 29.47 seconds.

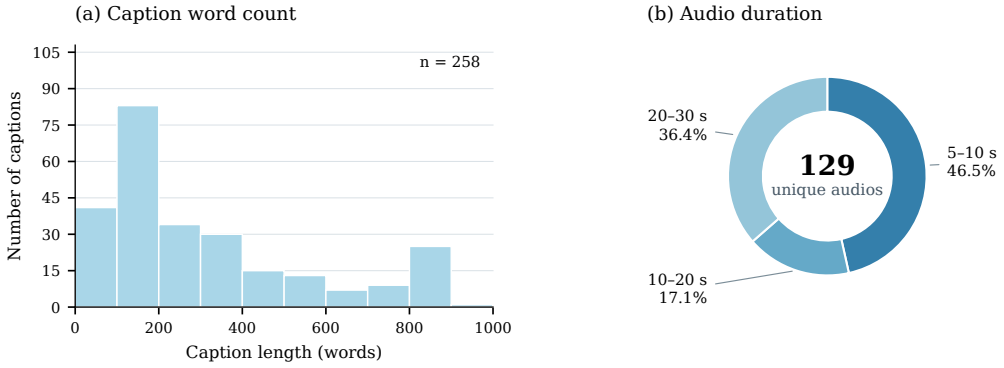


Figure 5: BRACE-Long statistics for the final 129-pair evaluation set. **(a)** Word counts for all 258 candidate captions, counted by whitespace with 100-word bins. **(b)** Original audio-file durations for 129 unique audios, counted once per audio.

A.3 Data Construction Prompts

Figure 6 consolidates five model descriptions into short, medium, and long anchors, prioritizing consensus and resolving conflicts conservatively. Using these textual anchors, GPT-5-mini evaluates four semantic components following X-ACE (Wang et al., 2024) and assigns an overall score (Figure 7). Component reasoning identifies matches, omissions, and hallucinations before assigning 0–100 scores; fluency penalties affect only the overall score. Teacher judgments depend on anchor quality.

The caption-shortening template (Figure 8) is shared by all five source systems and targets 12–30 words, allowing up to 40 when needed to preserve temporal or acoustic details. Figures 9–12 provide the four semantic perturbation templates. Figure 13 provides the shared data-corruption template, including all six instructions and examples. The placeholders are filled with the selected caption and, for data corruption, the selected instruction.

A.4 Training Data Format

Figure 14 shows an audio–caption input and its teacher-generated target. The student receives no anchors. Scores are normalized by dividing by 100 and serialized as five JSON objects, with overall first and score before reason, unlike the teacher's reasoning-first generation order.

High-Quality Reference Caption Generation Prompt

Generate high-quality regenerated golden captions for ONE audio clip using only the provided model-generated captions.

Your output must contain 3 length variants built from the same supported fact pool:
- gold_short (1-2 sentences)
- gold_mid (2-4 sentences)
- gold_long (4-7 sentences or one coherent paragraph)

Core policy:
- The model captions are the only evidence for the audio content.
- Do not invent unsupported details.
- Prefer consensus, but preserve salient details when they are clear and not contradicted.
- Resolve conflicts conservatively by choosing a less specific description.

Input schema (single clip JSON):

```
{
  "audio_id": "...",
  "sound": "...",
  "model_captions": {
    "af_next_captioner": "...",
    "gemini_3_flash_preview": "...",
    "qwen3_omni_captioner": "...",
    "moss_audio_8b_thinking": "...",
    "step_audio_r1_1": "..."
  }
}
```

Model guidance:
- Timestamp-aware models: af_next_captioner, moss_audio_8b_thinking.
  Use them for possible event order, timing, repetition, and transitions when the caption actually contains those details.
- Global description models: gemini_3_flash_preview, qwen3_omni_captioner, step_audio_r1_1.
  Use them for overall scene, ambience, and natural phrasing.
- Do not assume any model is always correct.

What to capture:
- Main sound events, sources, and actions.
- Acoustic attributes such as loudness, pitch, sharpness, muffling, or reverberation.
- Temporal structure such as sequence, repetition, continuity, interruption, or fading.
- Spatial cues such as nearby, distant, approaching, or moving.
- Scene context such as indoor/outdoor setting, ambience, background noise, and acoustic space.
- Layering, including foreground versus background sounds and overlapping events.

Regeneration rules:
1. Prioritize consensus. Events or details mentioned by multiple models are more reliable.
2. Resolve conflicts conservatively. Examples: ”truck” vs ”car” to ”vehicle”; ”man” vs ”person” to ”person”; ”shouting” vs ”speaking” to ”human voice” or ”speech.”
3. Avoid unreliable speech transcription. Do not include exact spoken words unless the same transcription is clearly supported by multiple captions and not contradicted. Otherwise, describe speech generally.
4. Avoid unsupported specificity. Do not add exact objects, locations, identities, emotions, causes, or actions unless clearly supported.
5. Preserve salient details. Keep important sound events, temporal order, repetition, background ambience, acoustic properties, and foreground/background relationships.
6. Write naturally. The output should be fluent natural-language audio captions, not a checklist or annotation table.

Length-specific output rules:
- gold_short:
  - 1-2 sentences
  - strongly compressed
  - focus on the most salient foreground sounds and at most one background layer

- gold_mid:
  - 2-4 sentences
  - include the main sound events, important ambience, and clear relationships

- gold_long:
  - 4-7 sentences or one coherent paragraph
  - richer but still evidence-grounded
  - no unsupported speculation and no duplicate listing

Output format (strict JSON only, no extra keys, no markdown):

```
{
  "audio_id": "...",
  "gold_short": "...",
  "gold_mid": "...",
  "gold_long": "..."
}
```

Input JSON:
{{INPUT_JSON}}

Figure 6: Prompt used to generate synthetic high-quality reference captions variants.

Candidate Caption Evaluation Prompt

You are an expert evaluator for Audio Captioning. Treat the reference caption set as a perfect description of the audio. Then evaluate how well the candidate caption aligns with it.

You MUST evaluate the captions across FOUR distinct components:
1. Sound Event: What sound events occur (e.g., siren, speech, footsteps, horning, barking)
2. Source: Who or what produces the sound (e.g., ambulance, man, woman, vehicle, dog)
3. Acoustic Attribute: Auditory characteristics of sound. How sound is perceived by humans (e.g., loud, high-pitched, continuous, playful, clear)
4. Temporal Relation: Timing and relationships between sound events (e.g., before, after, during, simultaneously, overlapping, background, foreground)

Evaluation Instructions:
- Evaluate EACH component independently.
- For each component:
    - First, write the reasoning by explicitly identifying:
        - what matches between candidate and reference
        - what is missing
        - what is incorrect or hallucinated
    - Then, assign an integer score from 0 to 100 based strictly on that reasoning.
    - The score MUST be a direct consequence of the reasoning.
    - Do NOT assign a score before completing the reasoning.
- Do not assign scores heuristically without explicit comparison to the reference caption set.
- Use the full 0-100 range when justified. Do not collapse scores into a few coarse bins such as only 0, 25, 50, 75, 100.
- A score of 100 means near-perfect alignment with the reference for that component.
- A score of 0 means the component is entirely incorrect, missing, or incompatible with the reference.
- After evaluating all four components:
    - Provide an overall reasoning and score that reflects the component-level evaluations
- Fluency Penalty (applies ONLY to overall score):
    - Reduce the overall score if the candidate contains severe repetition loops or broken/unfinished sentences that harm interpretability.
    - This penalty MUST NOT affect the four component scores.

Candidate set:
{candidate_statements}
Reference set:
{target_statements}

Output format (STRICT JSON):

```
{
  "sound_event": {"reason": "...", "score": 0},
  "source": {"reason": "...", "score": 0},
  "acoustic_attribute": {"reason": "...", "score": 0},
  "temporal_relation": {"reason": "...", "score": 0},
  "overall_reason": "...",
  "overall_score": 0
}
```

Figure 7: Prompt used to evaluate candidate captions against synthetic anchor captions across four components and assign an overall score.

Caption Shortening Prompt

System_prompt :
You rewrite audio captions. Shorten verbose captions while preserving sound events, sources, acoustic attributes, and temporal relations. Do not add events, infer new sources, or drop important audible details. Return only the requested JSON.

User_prompt :
Original caption:
{caption}

Rewrite it as one concise English audio caption. Aim for 12-30 words; use up to 40 words only if needed to preserve key temporal or acoustic details.

Structured output schema (API setting):

```
{
  "type": "object",
  "properties": {
    "short_caption": {
      "type": "string",
      "description":
        "A concise audio caption preserving the original sound meaning."
    }
  },
  "required": ["short_caption"],
  "additionalProperties": false
}
```

Format: `json_schema`; name: `short_audio_caption`; strict: `true`.

Figure 8: Prompt shared by all five source systems for shortening verbose captions, with the structured output schema supplied separately through the API.

Sound Event Perturbation Prompt

System_prompt :
You are a hallucination caption generator who can rewrite the caption into a
hallucinated caption that sounds plausible but is factually incorrect.
User_prompt :
You are given an audio caption. Your task is to modify only one Sound Event in the caption.
First, identify the Sound Events described in the caption. Replace exactly one existing Sound Event with a different Sound Event.
The modified caption must satisfy the following criteria:
1. Preserve all information except the selected Sound Event, including:
- Sources: who or what produces each sound
- Acoustic Attributes: auditory characteristics such as loudness, pitch, continuity, or clarity
- Temporal Relations: ordering, overlap, simultaneity, foreground/background, or other timing relationships between sound events
2. The replacement Sound Event must be significantly different from the original Sound Event.
3. The replacement must remain plausible for the same Source. Do not create an obvious semantic or physical contradiction.
4. Modify exactly one Sound Event. Do not add or remove sound events.
5. If no plausible but semantically distinct Sound Event replacement can be made while preserving the other information, return none.
6. Do not use near-synonyms or minor wording changes as replacements. If the best available replacement is only a near-synonym, return none.
—
Example 1:
Input: ”A woman laughs over repeated scraping.”
Output: ”A woman cries over repeated scraping.”
Explanation: Both “laughs” and “cries” are plausible human vocal actions, but they represent clearly different sound actions.
The caption remains coherent while introducing a meaningful semantic shift.
Example 2:
Input: ”A woman speaks over repeated scraping”
Output: ”A woman talks over repeated scraping.”
Explanation: “Speaks” and “talks” describe nearly the same sound action.
Near-synonym substitutions do not create a meaningful Sound Event hallucination, so this kind of weak negative is not allowed.
—
Input: {}
Output:
Only return the modified caption

Figure 9: Prompt for perturbing exactly one sound event while preserving the remaining caption information.

Source Perturbation Prompt

System_prompt :
You are a hallucination caption generator who can rewrite the caption into a hallucinated caption that sounds plausible but is factually incorrect.
User_prompt :
You are given an audio caption. Your task is to identify one noun from the caption and replace it with a new noun. The new noun must meet the following criteria:
1. Modify exactly one Source in the caption.
2. Preserve all other information, including:
- Sound Events: what sounds occur
- Acoustic Attributes: loudness, pitch, duration, distance, tone, etc.
- Temporal Relations: ordering, overlap, simultaneity, or other timing relationships between Sound Events
- Any other Sources not selected for modification
3. The replacement Source must be semantically distinct from the original Source. Do not use a synonym, subtype, or closely related entity.
4. The replacement Source must be able to plausibly produce the same Sound Event. Do not create a semantic or physical contradiction.
5. Do not add, remove, or modify any Sound Event, Acoustic Attribute, or Temporal Relation.
6. If no identifiable Source exists, or no plausible and sufficiently distinct replacement can be made while preserving the other information, return none.
Example 1:
Input: ”A woman speaks over repeated scraping”
Output: ”A man speaks over repeated scraping”
Explanation: Both ”woman” and ”man” can perform the action ”speaks over repeated scraping,” so the caption still makes logical sense. Replacing ”woman” with ”man” creates a meaningful semantic change in the context of the caption.
This is a good example (a positive case) because the substitution introduces semantic contrast while maintaining contextual coherence.
Example 2:
Input: ”A dog runs across the field.”
Output: ”A puppy runs across the field.”
Explanation: ”Dog” and ”puppy” both refer to similar animals, with ”puppy” being a younger version of a ”dog.” Although the substitution makes sense in the caption, the semantic difference between the two is relatively small.
This is a less effective example (a negative case) because the substituted noun is too close in meaning to the original and does not produce a significant semantic shift.
Input: {}
Output:
Only return the modified caption

Figure 10: Prompt for replacing exactly one sound source with a plausible, semantically distinct source.

Acoustic Attribute Perturbation Prompt

System_prompt :
You are a hallucination caption generator who can rewrite the caption into a hallucinated caption that sounds plausible but is factually incorrect.
User_prompt :
You are given an audio caption. Your task is to modify only one Acoustic Attribute associated with a Sound Event in the caption.
First, identify the Acoustic Attributes explicitly described in the caption. An Acoustic Attribute refers to an auditory characteristic of a Sound Event as perceived by humans, such as Volume, Pitch, Texture, Count, Tone, Timbre or other perceptual sound qualities.
Replace exactly one Acoustic Attribute with a different Acoustic Attribute.
1. Preserve all information except the selected Acoustic Attribute, including:
- Sound Events: what sounds occur
- Sources: who or what produces each sound
- Temporal Relations: ordering, overlap, simultaneity, foreground/background, or other timing relationships between sound events
2. The selected Acoustic Attribute may describe properties such as Volume, Pitch, Texture, Count, Tone.
3. Replace exactly one Acoustic Attribute with a perceptually and semantically distinct value, preferably within the same attribute dimension (e.g., high-pitched $\rightarrow$ low-pitched, faint $\rightarrow$ loud, distant $\rightarrow$ nearby).
4. Do not add, remove, or replace any Sound Event or Source, and do not modify unrelated Acoustic Attributes.
5. Avoid trivial or near-synonymous modifications that preserve nearly the same auditory meaning, such as loud $\rightarrow$ very loud, quiet $\rightarrow$ soft, or high-pitched $\rightarrow$ shrill.
—
Example 1:
Input: ”A faint engine hums continuously in the distance.”
Output: ”A loud engine hums continuously in the distance.”
Explanation: “Faint” and “loud” indicate clearly different volume levels. Only the Acoustic Attribute changes, creating a meaningful perceptual contrast.
Example 2:
Input: ”A loud engine is running.”
Output: ”A powerful engine is running.”
Explanation: “Loud” and “powerful” are too similar in auditory meaning. The change is plausible but not distinct enough.
—
Input: {}
Output:
Only return the modified caption

Figure 11: Prompt for perturbing exactly one acoustic attribute of an existing sound event.

Temporal Relation Perturbation Prompt

System_prompt :
You are a hallucination caption generator who can rewrite the caption into a hallucinated caption that sounds plausible but is factually incorrect.
User_prompt :
You are given an audio caption. Your task is to modify only the Temporal Relation between existing sound events in the caption.
First, identify the distinct Sound Events described in the caption.
If the caption contains fewer than two distinct Sound Events, return none.
If no explicit or inferable temporal relation exists between at least two Sound Events, return none.
Otherwise, generate a temporally hallucinated caption according to the following criteria:
1. Preserve all non-temporal information in the original caption, including:
- Sound Events: what sounds occur
- Sources: who or what produces the sounds
- Acoustic Attributes: auditory characteristics such as loudness, pitch, continuity, or clarity
2. Modify only the Temporal Relation between at least two existing sound events.
3. The modified temporal relation must be factually different from the original caption. Possible perturbations include:
- reversing event order: A before B $\rightarrow$ B before A
- changing before/after relations
- changing sequential events into simultaneous or overlapping events
- changing simultaneous or overlapping events into sequential events
—
Example 1:
Input: ”Men speaking then sneezing.”
Output: ”Men sneezing then speaking.”
Explanation:
The sound events, sources, and other non-temporal information remain unchanged. Only the temporal order between speaking and sneezing is reversed.
This is a good example because the hallucination is caused exclusively by a Temporal Relation perturbation.
Example 2:
Input: ”High pitched drilling followed by rustling.”
Output: ”High pitched drilling then rustling.”
Explanation: ”followed by” and ”then” express essentially the same temporal order. The wording changes, but the Temporal Relation does not, so this is not a meaningful hallucination.
Example 3:
Input: ”A dog barking loudly.”
Output: ”none”
Explanation:
Only one distinct Sound Event, barking, is present, so there is no Temporal Relation between multiple Sound Events to perturb.
—
Input: {}
Output:
Only return the modified caption.

Figure 12: Prompt for perturbing temporal relations between existing sound events.

Data Corruption Prompt

System_prompt :
You are a helpful assistant who can rewrite the caption based on the examples and requirements I have provided.
User_prompt :
I will provide you an audio caption.
Follow the instruction below and perturb the input caption.
I will give you an example.
Instruction: {}
Input caption: {}
Only output the perturbed caption without instructions in English!!!
—
Incomplete Caption. Rewrite the caption to make it incomplete by removing a key element while preserving the rest of the caption as much as possible.
Example:
Input: The teacher explained the homework assignment clearly to the students.
Output: The teacher explained the homework assignment clearly to.
Input : The teacher explained the homework assignment clearly to the students.
Output : The teacher explained the.
Repeated Event. Rewrite the caption to introduce unnecessary repetition of an event, making it sound redundant.
Example:
Input: The children started singing loudly.
Output: The children started singing, and then they sang loudly again.
Repeated Adverb. Rewrite the caption to add unnecessary repetition of an adverb, making it sound redundant.
Example:
Input: She walked slowly down the hall.
Output: She walked slowly, very slowly down the hall.
Missing Conjunction. Rewrite the caption by omitting a necessary conjunction, so the caption sounds slightly disconnected or incomplete.
Example:
Input: The car started moving, and the engine roared.
Output: The car started moving, the engine roared.
Missing Verb. Rewrite the caption by removing a main verb, making the caption feel like it lacks an action or event.
Example:
Input: The chef prepared the meal with great care.
Output: The chef the meal with great care.
Rewrite the Caption in Less Than Five Words. Rewrite the caption to condense it to five words or fewer while preserving its main idea as much as possible.
Example:
Input: The artist painted a beautiful landscape on a large canvas.
Output: The artist painted beautifully.

Figure 13: Shared data-corruption prompt with all six instructions and their examples. The first {} is replaced with the selected instruction, and the second with the input caption; all examples remain in the user message.

Training Data Format

User message

<audio>

Instruction: Judge the candidate caption based on audio-caption quality criteria and assign scores.

Evaluation axes:

- overall: holistic caption quality based on all relevant matches, omissions, and hallucinations.
- sound_event: what sound events occur, such as sirens, speech, footsteps, honking, or barking.
- source: who or what produces the sound, such as an ambulance, man, woman, vehicle, or dog.
- acoustic_attribute: auditory characteristics of sound and how it is perceived by humans, such as loud, high-pitched, continuous, playful, or clear.
- temporal_relation: timing and relationships between sound events, such as before, after, during, simultaneous, overlapping, background, and foreground.

Use normalized scores from 0.00 to 1.00 with exactly two decimal places.
0.00 means the caption does not describe the audio at all.
1.00 means the caption accurately and clearly describes the audio.

Return strict JSON only.
Include exactly these top-level keys in this order: overall, sound_event, source, acoustic_attribute, temporal_relation.
For every object, write "score" before "reason".
Do not include any extra keys.

Return exactly this JSON schema:

```
{
  "overall": {"score": 0.00, "reason": "..."},
  "sound_event": {"score": 0.00, "reason": "..."},
  "source": {"score": 0.00, "reason": "..."},
  "acoustic_attribute": {"score": 0.00, "reason": "..."},
  "temporal_relation": {"score": 0.00, "reason": "..."}
}
```

Candidate caption: The audio consists of a dog barking three times in rapid succession. The barks are sharp, clear, and relatively high-pitched, characteristic of a medium-sized dog. There is no other background noise, making the barking sound very distinct and isolated.

Assistant target

```
{
  "overall": {"score": 0.96, "reason": "…"},
  "sound_event": {"score": 1.00, "reason": "…"},
  "source": {"score": 0.90, "reason": "…"},
  "acoustic_attribute": {"score": 0.95, "reason": "…"},
  "temporal_relation": {"score": 1.00, "reason": "…"}
}
```

Figure 14: Training example with the full user instruction and assistant target.

A.5 BENCHMARK EVALUATION PROMPTS

Figures A.5 and 16 show fixed, reference-free templates used across BRACE (Guo et al., 2025), BRACE-Long, RELATE (Kanamori et al., 2025), and PAM (Deshmukh et al., 2024). Baselines use an audio-adapted FLEUR prompt (Lee et al., 2024b); our judge uses its training instruction and five-output JSON format. FLEUR smoothing uses digit-token probabilities separately from the generated score; Gemini results use raw scores because these probabilities were unavailable.

FLEUR Evaluation Prompt

Your task is to evaluate and rate the caption on a scale of 0.0 to 1.0 based on the given Grading Criteria. (Print Real Number Score ONLY)

Grading Criteria:

0.0: The caption does not describe the audio at all.
1.0: The caption accurately and clearly describes the audio.

Caption: {pred_caption}

Score(Choose a rating from 0.0 to 1.0):

Figure 15: FLEUR evaluation prompt with the original one-decimal score scale.

Audio-Compass Evaluation Prompt

Instruction: Judge the candidate caption based on audio-caption quality criteria and assign scores.

Evaluation axes:
- overall: holistic caption quality based on all relevant matches, omissions, and hallucinations.
- sound_event: what sound events occur, such as sirens, speech, footsteps, honking, or barking.
- source: who or what produces the sound, such as an ambulance, man, woman, vehicle, or dog.
- acoustic_attribute: auditory characteristics of sound and how it is perceived by humans, such as loud, high-pitched, continuous, playful, or clear.
- temporal_relation: timing and relationships between sound events, such as before, after, during, simultaneous, overlapping, background, and foreground.

Use normalized scores from 0.00 to 1.00 with exactly two decimal places.
0.00 means the caption does not describe the audio at all.
1.00 means the caption accurately and clearly describes the audio.

Return strict JSON only.
Include exactly these top-level keys in this order: overall, sound_event, source, acoustic_attribute, temporal_relation.
For every object, write "score" before "reason".
Do not include any extra keys.

Return exactly this JSON schema:

```
{
  "overall": {"score": 0.00, "reason": "..."},
  "sound_event": {"score": 0.00, "reason": "..."},
  "source": {"score": 0.00, "reason": "..."},
  "acoustic_attribute": {"score": 0.00, "reason": "..."},
  "temporal_relation": {"score": 0.00, "reason": "..."}
}
```

Candidate caption: {pred_caption}

Figure 16: Fixed evaluation template for Audio-Compass, identical to the training instruction in Figure 14.